Evaluating Bilingual Lexicon Induction without Lexical Data

Michaela Denisová and **Pavel Rychlý**
Faculty of Informatics
Masaryk University
Brno, CZ, 602 00
449884@mail.muni.cz, pary@fi.muni.cz

Abstract

Bilingual Lexicon Induction (BLI) is a fundamental task in cross-lingual word embedding (CWE) evaluation, aimed at retrieving word translations from monolingual corpora in two languages. Despite the task's central role, existing evaluation datasets based on lexical data often contain biases such as a lack of morphological diversity, frequency skew, semantic leakage, and overrepresentation of proper names, which undermine the validity of reported performance. In this paper, we propose a novel, language-agnostic evaluation methodology that entirely eliminates the dependency on lexical data. By training two sets of monolingual word embeddings (MWEs) using identical data and algorithms but with different weight initialisations, we enable the assessment on the BLI task without being affected by the quality of the evaluation dataset. We evaluate three baseline CWE models and analyse the impact of key hyperparameters. Our results provide a more reliable and bias-free perspective on CWE models' performance.

1 Introduction

Bilingual Lexicon Induction (BLI) is an intrinsic evaluation task designed to identify and extract translations of individual words. This task has been a widely adopted method for evaluating cross-lingual embeddings (CWEs), which aim to align two (or more) sets of independently trained monolingual word embeddings (MWEs) into a shared cross-lingual space, where semantically similar words are represented by closely aligned vectors (Ruder et al., 2019).

Given this characteristic, they have shown significant advantages in numerous NLP applications, such as machine translation (Duan et al., 2020; Zhou et al., 2021; Wang et al., 2022), cross-lingual information retrieval (Vulić and Moens, 2015), language acquisition and learning (Yuan et al., 2020).

In the BLI task, the method aims to generate a list of target words for each source word, ranking them based on the cosine similarities between their respective embeddings. Afterwards, top-*k* target words for each source word are selected, and the word pairs are compared to the evaluation dataset (Ruder et al., 2019).

Over the years, a large dataset of 110 bilingual dictionaries MUSE (Conneau et al., 2017) published along with the strong eponymous baseline model has become the standard benchmark for the BLI task in many papers. (Joulin et al., 2018; Severini et al., 2022; Cao et al., 2023; Ding et al., 2024, 2025) Despite their popularity, the MUSE datasets have been subject to criticism concerning their reliability in reflecting model performance.

Kementchedjhieva et al. (2019) and Laville et al. (2022) reported a high proportion of proper names and graphically identical or similar word pairs, over 30% and 40% respectively. Czarnowska et al. (2019) further noted that the datasets are skewed toward high-frequency words, suffer from semantic leakage between training and evaluation sets, and lack morphological diversity.

While prior work has proposed alternative evaluation datasets to address the limitations of MUSE, many of these efforts still fall short in reliability. For instance, some include only one-to-one word pairs (Glavaš et al., 2019), neglecting polysemy. Others compile their datasets automatically (Glavaš et al., 2019; Vulić et al., 2019), inheriting many of the same issues as MUSE.

Evaluation based on lexical data introduces biases, including missing valid translations, frequency skew, semantic leakage, and limited morphological or lexical diversity. A core issue lies in the mismatch between evaluation data, typically base forms or limited variants, and embedding spaces trained on inflected forms. As a result, valid translations like German *läuft* for English *run* are

Proceedings of Recent Advances in Natural Language Processing,pages 275–282
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-034

marked incorrect if only *laufen* appears in the evaluation dataset.

These inconsistencies also impact the alignment process. A single lemma can correspond to multiple inflected word forms, e.g., *light* may translate to *licht* (noun), *leicht* (adjective), or *beleuchten* (verb) in German. Such variation introduces ambiguity and noise into the learned mappings, particularly in morphologically rich languages, where the number of possible word forms per lemma is large. Yet, it is often unclear whether poor performance stems from data bias or algorithmic limitations.

Consequently, it hinders our ability to correctly interpret the results, make accurate comparisons between the proposed solutions, and monitor the progress reliably. Moreover, addressing these biases is both time-consuming and challenging, often requiring specialised linguistic expertise and careful manual intervention.

Motivated by these insights, we propose a novel evaluation methodology for assessing the quality of aligned embeddings on the BLI task without relying on any lexical data. Specifically, we train two sets of MWEs using the same data and algorithm while creating different weight initialisations for each embedding space. The aim is to ensure that both embedding spaces capture the same underlying distributional patterns, with differences arising only from stochastic variation. This allows us to determine the upper bound of CWE models' performance, evaluating how effectively algorithms can perform under ideal data conditions.

We apply this framework to evaluate three baseline CWE models. Additionally, we investigate the impact of key hyperparameters, such as embedding dimension, number of epochs, size of the seed lexicon, and word frequency, on the models' performance.

Our contribution is threefold.

1. We introduce a novel approach for evaluating aligned embeddings in the BLI task that does not rely on lexical data and avoids common biases found in existing datasets.

2. We provide a reliable and accurate evaluation of three baseline CWE models with different levels of supervision, independent of known issues in standard evaluation datasets.

3. We systematically investigate how various hyperparameters affect the performance of baseline CWE models.

2 Background

Advancements in BLI have largely been driven by progress in CWE methods, where BLI serves as a key intrinsic evaluation task. The pioneering work introducing the embedding-based method evaluated on the BLI task was proposed by Mikolov et al. (2013), spawning an immense number of articles continuing in the research. These proposed methods ranged from classical, state-of-the-art baseline models such as MUSE (Conneau et al., 2017), VECMAP (Artetxe et al., 2018b,a), RCLS (Joulin et al., 2018) to current endeavours in integrating contextual embeddings and Large Language Models (LLMs). (Vulić et al., 2023; Li et al., 2023; Hu and Xu, 2024)

Despite the abundance of proposed solutions, the evaluation has received limited attention (Laville et al., 2022). One of the first comprehensive evaluations was conducted by Glavaš et al. (2019), who systematically compared projection-based CWE models across both intrinsic and extrinsic tasks. Their findings showed that optimising CWE models solely for BLI can result in degraded downstream performance. Moreover, the authors constructed standardised datasets across 28 language pairs using frequent English words and their translations via Google Translate.

Later studies shifted focus toward analysing the standard evaluation datasets MUSE, highlighting several of its limitations. Kementchedjhieva et al. (2019) conducted a study revealing that approximately a quarter of the word pairs consist of proper names (e.g., *Barack Obama*, *Skype*), which are often graphically identical across languages. As a result, they advocated for the adoption of more reliable evaluation methodologies or for performing an evaluation with rigorous error analysis.

Another analysis provided by Czarnowska et al. (2019) showed that all word pairs in the MUSE datasets are drawn from the 10K most frequent words in each language. They also discovered that these datasets suffer from semantic leakage, where it is common for a word to appear in both the training and evaluation datasets in different inflected forms. Finally, they mentioned that the MUSE datasets lack morphological diversity, where most words occur in only one inflected form. As a solution, they compiled morphologically complete datasets for 5 Slavic and 5 Romance languages.

The most recent study conducted by Laville et al. (2022) pointed out that many MUSE datasets contain over 30% identical word pairs, such as proper names (*Frederico*, *Brian*), brands (*android*), geographical entities (*Gelsenkirchen*, *Nebraska*), or words from other languages, mostly English (*freedom*, *musica*). Additionally, they revealed that, on average, over 40% of word pairs in the datasets are graphically similar.

There have been several efforts to address these limitations. Izbicki (2022) introduced manually annotated datasets across 298 languages in combination with English. They focused on a uniformly distributed part of speech in each dataset to make the results as comparable as possible. Other attempts automatically compiled datasets for diverse language combinations to mitigate English-centric bias. (Vulić et al., 2019; Anastasopoulos and Neubig, 2020)

3 Experimental Setup

In this section, we outline the key components of the experiments that were conducted.

3.1 Monolingual Embeddings

We train the fastText algorithm (Bojanowski et al., 2017) with dimensions of 300, 100, 50, and 20, each trained for 1, 3 and 5 epochs. For every configuration, we generate two embedding spaces, A and B. In space B, we modify the vocabulary by prefixing each letter of every word with ”x” (e.g., *apple* becomes *xaxpxpxlxe*). We trim all vocabularies to the 300K most frequent words.

3.2 Cross-lingual Embeddings

To retrieve aligned MWEs, we utilised three state-of-the-art CWE frameworks, RCLS in a supervised mode and MUSE and VECMAP (VM) in a supervised (MUSE-S, VM-S), unsupervised (MUSE-U, VM-U) mode and a mode that relies on identical strings (MUSE-I, VM-I).

The default settings closely followed the MUSE training described in Conneau et al. (2017), VM-S and VM-I in Artetxe et al. (2018a), VM-U settings in Artetxe et al. (2018b), and RCLS settings in Joulin et al. (2018).

3.3 Data

We train our MWEs using the fastText algorithm on a corpus derived from the Czech Wikipedia dump [1], containing approximately 170 million tokens. We preprocess the raw Wikipedia data using Matt Mahoney’s normalisation Perl script. [2]

Group	Max	Min
1	80,772,389	19,153
2	19,134	5,840
3	5,836	2,233
4	2,231	715
5	712	6

Table 1: Minimum and maximum frequency values by group.

To train CWEs in supervised mode, we randomly sample 5K words from the MWE vocabulary and create word pairs by pairing each word with a transformed version, in which every letter is prefixed with ”x”.

To evaluate CWEs, we split the MWE vocabulary into five frequency groups (36.5K words each). Table 1 shows their frequency ranges. From each group, we randomly sample 2K words (5.5%), totalling 10K evaluation words. For each evaluation word, we retrieve its top-1 nearest neighbour from the target space, strip the inserted ”x” characters, and check whether the retrieved word matches the original word.

3.4 Metric

We report Precision@k (P@k), which is denoted by the following formula:

$$P = \frac{TP}{(TP + FP)} \tag{1}$$

, where TP (true positives) denotes the number of correctly retrieved target-word candidates that match the target words from the evaluation dataset, and FP (false positives) represents the number of incorrect target-word candidates retrieved by the system. And, P@k evaluates the proportion of TPs among the top-k predicted candidates for each source word. In this paper, k = 1.

4 Evaluation

This section presents the main results, organised into four parts, each examining the impact of a key hyperparameter.

4.1 Dimension Impact

Overall results are displayed in Table 2. Most models achieved a remarkable performance of 100%,

[1] https://dumps.wikimedia.org/cswiki/20250301/

[2] http://mattmahoney.net/dc/textdata.html

especially when using 100- and 300-dimensional embeddings across all epochs. An exception is MUSE-I, which failed to align any embeddings due to its reliance on identical strings, absent in our setup. Additionally, performance declined slightly with 50 and 20-dimensional embeddings, within a margin of around 0.1% to 2%. In contrast, RCLS does not consistently benefit from higher embedding dimensions (e.g., 99.4% at DIM 20 vs. 85.6% at DIM 100), and yields the weakest results among all models, reaching only 85.6% P@1 when using 100-dimensional embeddings trained for one epoch.

In the next phase, we investigated the effect of systematic dimensionality ablation on models' performance. For each set of pre-trained MWEs (20-, 50-, 100-, and 300-dimensional), we gradually truncated the embeddings by retaining only a subset of dimensions. Specifically, from the 100-dimensional embeddings, we aligned versions with the last 10, 20, 40, 50, 60, 70, 80, and 90 dimensions removed. For the 300-dimensional embeddings, we retained the first 150, 100, 80, 60, 50, and 20 dimensions. In the case of the 50- and 20-dimensional embeddings, we evaluated a version preserving the first 20 and 10 dimensions. This setup allows us to explore how many dimensions can be removed to still carry meaningful information and maintain strong performance. The results are illustrated in Fig. 1, 2, and 3.

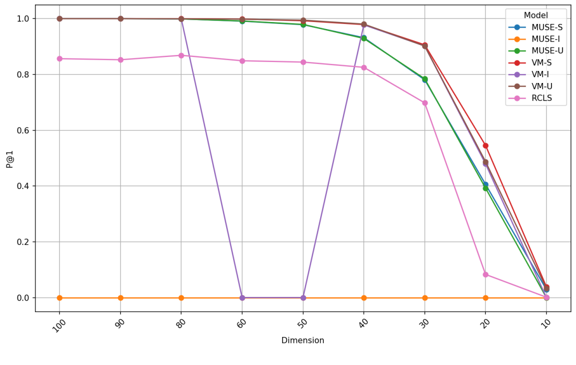

Figure 1: P@1 performance across models after reducing dimensions from 100-dimensional MWEs trained for one epoch.

When looking at Fig. 1 and 2, we can observe that the majority of models maintain a high performance of 100% until the dimensions fall to approximately 40. On the other hand, RCSLS achieves the worst performance and shows a more pronounced decline at lower dimensions. VM-I also fails as dimensionality is reduced, collapsing at 80 dimensions in the 300-dimensional setting and experiencing a sharp performance drop to 0% at 60 and 50 dimensions in the 100-dimensional setting, indicating that dimensions are not uniformly informative.

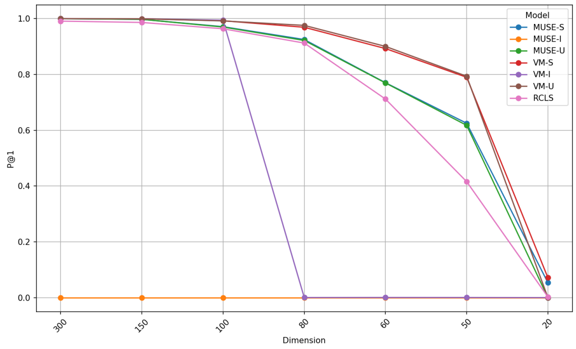

Figure 2: P@1 performance across models after reducing dimensions from 300-dimensional MWEs trained for one epoch.

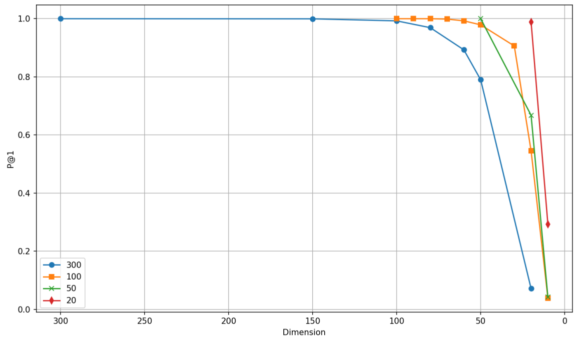

Figure 3: P@1 comparisons between reduced dimensions from 300, 100, 50, and 20-dimensional MWEs trained for one epoch and aligned using model VM-S.

On top of that, Fig. 3 shows that 100-, 50- and 20-dimensional MWEs retain meaningful information more effectively under dimensionality constraints compared to 300-dimensional embeddings. This suggests that fewer dimensions are sufficient to achieve high performance in CWE models, while 100-dimensional embeddings provide an efficient compromise between size and alignment quality.

Due to the significant performance fluctuations observed with the VM-I model at 50 dimensions, we further investigated whether alternative dimensionality selection strategies could yield improved results. Specifically, we evaluated two configurations: one retaining the last 50 dimensions instead of the first, and a new embedding constructed by selecting every second dimension from the original embeddings. Table 3 suggests that alternative selection strategies can boost the models' performance to nearly 100%.

	MUSE-S	MUSE-I	MUSE-U	VM-S	VM-I	VM-U	RCLS
DIM 20 EP 1	98.0	0.0	98.2	98.8	98.8	98.8	92.8
DIM 20 EP 3	99.9	0.0	99.8	99.9	99.9	99.9	99.4
DIM 20 EP 5	97.8	0.0	97.9	99.1	0.0	99.1	97.1
DIM 50 EP 1	99.9	0.0	99.9	100	100	100	99.1
DIM 50 EP 3	99.9	0.0	99.9	100	99.9	99.9	99.8
DIM 50 EP 5	99.9	0.0	99.9	99.9	100	100	99.8
DIM 100 EP 1	100	0.0	100	100	100	100	85.6
DIM 100 EP 3	100	0.0	100	100	100	100	99.9
DIM 100 EP 5	100	0.0	100	100	100	100	99.9
DIM 300 EP 1	100	0.0	100	100	100	100	99.1
DIM 300 EP 3	100	0.0	100	100	100	100	99.9
DIM 300 EP 5	100	0.0	100	100	100	100	99.9

Table 2: P@1 (%) across all the models with different dimension (DIM) and epoch (EP) configurations.

Embedding	P@1 (%)
LAST E1	99.71
LAST E5	99.44
SKIP E1	99.63
SKIP E5	99.62

Table 3: P@1 performance of VM-I using alternative 50-dimensional subsets: last 50 dimensions (LAST) and every second dimension (SKIP) across 1 and 5 training epochs (E).

4.2 Epoch Impact

Fig. 2 and 5 outline the dimensionality reduction performance when using 300-dimensional MWEs trained for one and five epochs, respectively. Figure 4 shows how training epochs affect the performance of the MUSE-S model under dimensionality reduction from 300 dimensions.

Overall, the three-epoch models achieve consistently the best performance, offering a trade-off between training duration and alignment quality. On top of that, the model trained for a single epoch outperforms the five-epoch version across all reduced dimensions. The exception is the model RCLS, which yields better results with a higher number of epochs involved. These results suggest that while shorter training helps preserve performance under dimensionality reduction, more epochs may still benefit weaker models by enhancing their overall alignment performance.

4.3 Seed Lexicon Impact

In this set of experiments, we compared supervised models with their unsupervised counterparts by progressively reducing the size of the seed lexicon from 5K seeds to 2.5K, 1K, 500, 100, and 20 seeds. The objective was to determine the minimum number of seeds required for the supervised models to outperform the unsupervised ones. To this end, we evaluated MUSE and VM in both supervised and unsupervised modes, training them using MWEs with 20 dimensions (reduced from a 100-dimensional space) and 50 dimensions (reduced from a 300-dimensional space). The results are visualised in Fig. 6 and 7.

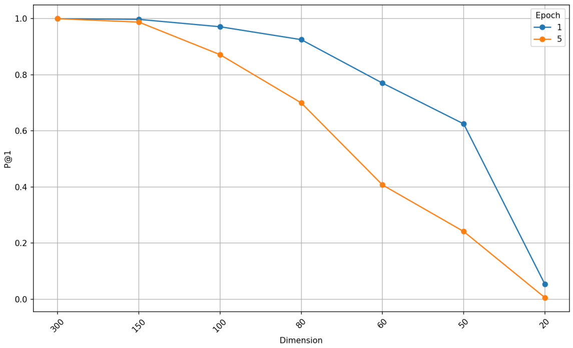

Figure 4: Comparison of the MUSE-S model's performance while using MWEs trained for one and five epochs.

In both figures, VM-S trained with one epoch MWEs consistently outperforms most other models, maintaining relatively high performance until the seed lexicon size falls below 100, at which point all models experience a sharp decline. In contrast, Fig. 6 shows that MUSE-S maintains more stable but lower performance, largely unaffected by seed lexicon size in higher ranges, but falls rapidly with minimal supervision. Notably, in Fig. 7, we can observe that VM-U surpasses all its supervised counterparts.

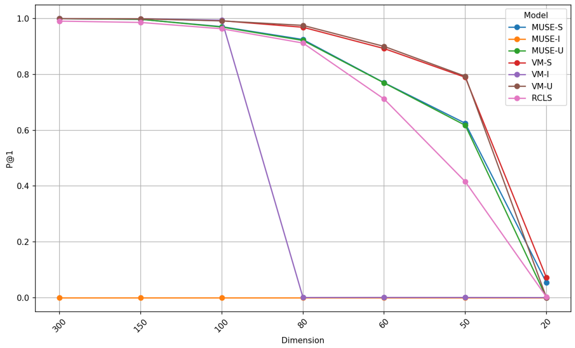

Figure 5: P@1 performance across models after reducing dimensions from 300-dimensional MWEs trained for five epochs.

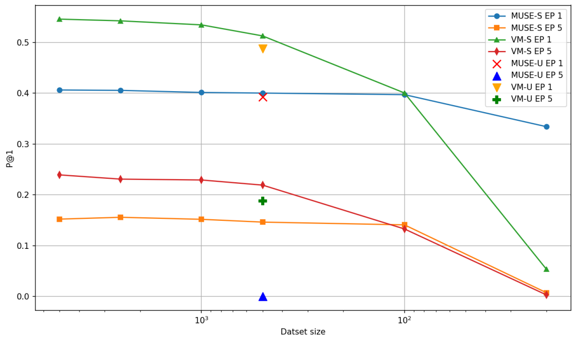

Figure 6: P@1 performance of supervised and unsupervised MUSE and VM models across decreasing seed lexicon sizes, using 20-dimensional embeddings reduced from 100 trained for 1 and 5 epochs (EP).

4.4 Frequency Impact

Finally, we examined how the models perform for words from different frequency groups. Fig. 8 presents the model VM-S and its P@1 performance distribution across five frequency groups at various embedding dimensions.

Overall, the model performs in all frequency groups nearly identically, especially at higher and very low dimensions, with P@1 scores varying within a margin of less than 1%, indicating that word frequency has little impact on alignment quality. At 30 and 20 dimensions, modest differences emerge, with mid-frequency groups slightly outperforming the highest and lowest ones by up to 4% to 6%.

5 Conclusion

In this paper, we introduced a novel evaluation methodology for assessing CWE models on the BLI task without relying on lexical data. This approach enables a bias-free and reliable evaluation, independent of the limitations found in standard datasets. We evaluated three baseline CWE models under varying levels of supervision and examined the effects of key hyperparameters, including embedding dimensionality, number of training epochs, seed lexicon size, and word frequency.

Our findings reveal that 100-dimensional embeddings offer an effective trade-off between compactness and alignment quality. Embeddings trained with fewer epochs generally yield better performance, though additional training can benefit models with weaker initial results, such as RCLS. Among all evaluated models, VM-S proved the most robust, maintaining high precision even under low supervision. Finally, performance remained largely consistent across all word frequency groups, highlighting the general reliability of the models under controlled evaluation conditions.

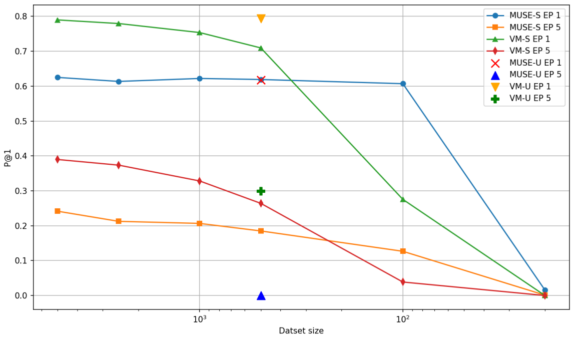

Figure 7: P@1 performance of supervised and unsupervised MUSE and VM models across decreasing seed lexicon sizes, using 50-dimensional embeddings reduced from 300 trained for 1 and 5 epochs (EP).

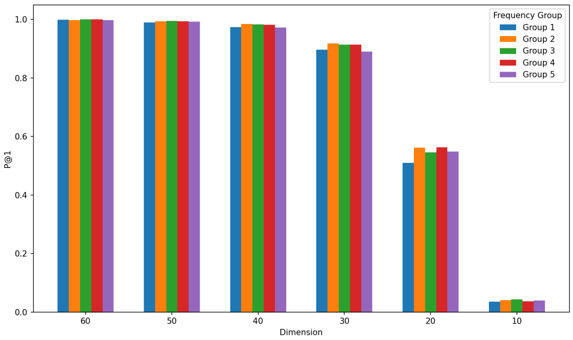

Figure 8: P@1 across frequency groups for the VM-S model trained with different embedding dimensions.

References

Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–

8679, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Hailong Cao, Tiejun Zhao, Weixuan Wang, and Wei Peng. 2023. Bilingual word embedding fusion for robust unsupervised bilingual lexicon induction. *Information Fusion*, 97:101818.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv'e J'egou. 2017. Word translation without parallel data. *ArXiv*, abs/1710.04087.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 974–983, Hong Kong, China. Association for Computational Linguistics.

Qiuyu Ding, Hailong Cao, Zihao Feng, Muyun Yang, and Tiejun Zhao. 2025. Enhancing bilingual lexicon induction via harnessing polysemous words. *Neurocomputing*, 611:128682.

Qiuyu Ding, Hailong Cao, and Tiejun Zhao. 2024. Enhancing bilingual lexicon induction via bi-directional translation pair retrieving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17898–17906.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, Online. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Ling Hu and Yuemei Xu. 2024. DM-BLI: Dynamic multiple subspaces alignment for unsupervised bilingual lexicon induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2041–2052. Association for Computational Linguistics.

Mike Izbicki. 2022. Aligning word vectors on low-resource languages with Wiktionary. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 107–117. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Martin Laville, Emmanuel Morin, and Phillippe Langlais. 2022. About evaluating bilingual lexicon induction. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 8–14. European Language Resources Association.

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9577–9599. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *The Journal of Artificial Intelligence Research*, 65:569–631.

Silvia Severini, Viktor Hangya, Masoud Jalili Sabet, Alexander Fraser, and Hinrich Schütze. 2022. Don't forget cheap training signals before building unsupervised bilingual word embeddings. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 15–22. European Language Resources Association.

Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti, and Anna Korhonen. 2023. Probing cross-lingual lexical knowledge from multilingual sentence encoders. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2089–2105. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877. Association for Computational Linguistics.

Xingdi Yuan, Jie Fu, Marc-Alexandre Côté, Yi Tay, Chris Pal, and Adam Trischler. 2020. Interactive machine comprehension with information seeking agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2325–2338, Online. Association for Computational Linguistics.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834. Association for Computational Linguistics.